Latin American Data in Science, v.1, n.3
datainscience.com.br
ISSN: 2763-9290

The Fitnome Catalog: a resource for physical exercise genetics data mining.

Recebido: 25/10/2021 | Aceito: 16/11/2021 | Publicado: 05/12/2021
https://doi.org/10.53805/lads.v1i3.32

Christina P. S. Martin[1], Stela M. da S. Felipe[2], Juliana O. Alves[3], Raquel M. de Freitas[2], Luis Henrique P.dos Santos[2], Adriano César C. Loureiro[2], Paula M. Soares[2], Vânia M. Ceccatto[2*]

ABSTRACT

Physical exercise (PE) in regularity is a well-characterized non-pharmaceutical intervention for good health and welfare. Molecular mechanisms regulated in response to PE can be scrutinized, with molecular biology, genomics, transcriptomics, and bioinformatics being inserted into exercise physiology studies. From a biotechnological perspective, omic datasets about physical exercise gene expression help identify phenotypic, genetic variance for different physical training phenotypes. Extensive lists of genes regulated by PE were dispersed within the literature, and the Fitnome Catalog (FitC) was created to reach some systematization of this information. Manual and online text-mining tools generated this dataset in PE human gene expression articles (2003-2014) with microarray, RNA-Seq, RT-PCR, and genotyping methods. Spreadsheets were developed with information on exercise protocol, experimental design, gender, age, number of individuals, analytical approach, gene ID, fold change and statistical data, and genetic architecture, encompassing 21 columns. The produced dataset (with 5,147 genes and 101,343 data points) provides experimental design, gene expression information, gene attributes, and references. Functional categorization of the FitC dataset and standardized information on PE-expressed genes were presented.

Keywords: Physical exercise, gene expression, differentially expressed genes (DEG), genetic dataset, 'omic analysis.

PRIOR PUBLICATIONS

PACHECO, C. et al. A compendium of physical exercise-related human genes: an 'omic scale analysis. Biology of Sports. 35(1):3-11. Mar, 2018. DOI: https://doi.org/10.5114/biolsport.2018.70746.

PACHECO, C. et al. Regulação gênica da via AMPK pelo exercício físico: revisão sistemática e análise in silico. Rev Bras Med Esporte 23 (04), Jul-Aug 2017, DOI: 10.1590/1517-869220172304169935.

[1] Universidade de Brasília - (UnB), Brasília, DF, Brazil.
[2] Universidade Estadual do Ceará (UECE) – Instituto Superior de Ciências Biomédicas – ISCB – Laboratório de Bioquímica e Expressão Gênica – LABIEX/UECE Fortaleza, CE, Brazil. vania.ceccatto@uece.br.
[3] Universidade Estácio do Ceará (ESTACIO) – Departamento de Educação Física, Fortaleza, CE, Brazil.

DATA IMPORTANCE

- Physical Exercise (PE) induces gene expression regulation, leading to human regulatory systems. It is a non-pharmaceutical intervention for type 2 diabetes (COLBERG et al., 2010), heart disease (SEO et al., 2020), anxiety (STRÖHLE, 2009), and Alzheimer's (VASCONCELOS-FILHO et al., 2021);
- Molecular mechanisms underlying the acute adaptations to PE can be scrutinized using molecular biology, bringing transcriptomics into exercise physiology studies (GOMES et al., 2020). DEGs extensive lists were dispersed within the literature as supplementary tables with various formats;
- A new challenge for molecular exercise physiology researchers is how to deal with "big data": understanding, analyzing, and applying the generated knowledge in the field (VOLTARELLI; FERNANDES; BRUM, 2020). Research with post-genomic analysis using 'omics data brings advances to PE studies;
- The Fitnome Catalog is a step in this "big data" analysis: the systematization of information from the scientific literature. The dataset (with 5,147 genes and 101,343 data points) provides experimental design data, gene expression information, gene attributes, and references (PACHECO et al., 2018).

MATERIALS AND METHODS

Data collection and Fitnome Catalog formatting

Public literature repositories PMC, PubMed, and Google Scholar were searched for PE differential gene expression literature based on the microarray, RNA-Seq, RT-PCR, and genotyping methods. After text-mining through 122 records, 75 papers were worked on, and 58, from the 2003 to 2014 period, were selected for the study. The text-mining approach (FONTAINE et al., 2011) was used to search for additional research papers linking human genes to physical exercise. Additional references supplied by these works were also investigated.

The gene lists were collected from these papers, including supplementary material. Data from the produced spreadsheets were merged, and duplicates were removed using Excel's decision tools. The following additional data were collected from the studies: exercise protocol, experimental design (acute/chronic), gender, age range, number of individuals tested, analytical method, and reference. Genetic architecture information was collected with *Ensembl's BioMart MartView* tool *(http://biomart.org)*, generating a spreadsheet with gene attributes (KINSELLA et al., 2011). Data from the produced spreadsheets were carefully merged using Excel's decision tools. The resulting spreadsheet contains 5,147 genes and 101,343 cells with information on exercise protocol, experimental design, gender, age, number of individuals, analytical method, fold change and statistical data.

Functional categories of the FitC gene set

The collected gene list comprehends extensive material for further analyses. All bioinformatics analyses performed by us were based on Human Genome Assembly GRCh38, and the number of annotated genes was obtained by browsing the Ensembl database on 07/04/2015 (MAGLOTT et al., 2007). The actualization of the gene data and identification (Column 1 – Supplementary material) is possible by a link for the Ensembl Gene ID to the origin database site. The reference list was all mapped *Entrez gene* IDs from the selected platform genome. This reference list mapped 61,506 *Entrez gene IDs (ncbi.nlm.nih.gov/gene)*. 14,809 IDs annotated to the selected functional categories that are used as the reference for the enrichment analysis. The list contained 5,144 user IDs in which 4,840 user IDs are unambiguously mapped to 4,840 unique Entrez gene IDs, and 304 user IDs can not be mapped to any Entrez gene ID. The *GO Slim* (*GO = Gene Ontology*) *(ebi.ac.uk/)*

(ORTON et al., 2016) summary is based upon the 4,840 unique Entrez gene IDs. *GO Slim* was a list of selected terms creating high-level summaries of an area of biology (canonical metabolic pathways) based on selected *GO terms*.

The functional categorization of the FitC dataset was performed by gene enrichment analysis on the gene-set (Fig. 1) by the Over Representation Analysis (ORA) method (BOYLE et al., 2004) based on the online tool *WEB-based Gene SeT AnaLysis Toolkit (http://www.webgestalt.org/)* (LIAO et al., 2019). The used parameters were: Organism: *Homo sapiens*, Enrichment Categories: gene ontology biological process nonredundant. Figure 2 shows the FitC gene hierarchy on Directed Acyclic Graph (DAG) based showing canonical metabolic pathways. Among 4,840 unique Entrez gene IDs, 3,935 IDs were annotated to the selected functional categories and the reference list used for the enrichment analysis. Statistics were done by *False Discovery Rate* (FDR) method by Benjamini-Hochberg procedure, with FDR ≤ 0.05. The Benjamini–Hochberg method (BENJAMINI; HOCHBERG, 1995) controls the FDR using sequential modified Bonferroni correction for multiple hypothesis testing. Based on the above parameters, 12 categories were identified (Fig. 1A) as enriched biological processes with a strong representation of the following processes: "biological regulation," "metabolic process," and "response to a stimulus." When evaluating molecular function (Fig. 1B), the gene set was well represented in 19 categories, with the strongest links to "protein binding," "ion binding," and "nucleic acid binding." FitC gene ontology (DAG) (Fig. 2) presents GO statistically significant terms. These terms are significantly correlated with PE issues about energy and muscle metabolism like tissue migration, muscle system process, muscle tissue development, generation of precursor metabolites and energy, extracellular structure organization, muscle cell differentiation, etc.

DATA DESCRIPTION

The dataset presented here is a spreadsheet entitled: "Fitnome Catalog" That file is composed of two tabs. The first is called "The Fitnome Catalog" with 5,149 lines x 21 columns, and the second is called "Fitnome Catalog References" with 59 lines x one column (Supplemental Material).

Dataset

The Fitnome Catalog columns fields:

A - Ensembl Gene ID: gene label identifier in the *Ensembl database (ensembl.org)*. The identifiers were used to be unambiguous and consistent across Ensembl releases. Every Ensembl Gene ID is linked to the origin site record.

B - Gene symbol: gene symbol label or tag.

C - Description: full gene name.

D - Gene type: protein-coding, long noncoding, short noncoding, pseudogenes, and not yet classified genes.

E - Chromosome: chromosome number in the human nuclear genome.

F - Band: genetic band or portion of the chromosomal location.

G - Strand: gene location on DNA strand (1= positive or template strand; -1=negative or non-template strand).

H - Gene Start (bp): initial nucleotide sequence position on the chromosome.

I - Gene End (bp): final nucleotide sequence position on the chromosome.

J - Linked to what type of exercise: significant type of physical exercise performed in the study (categories: endurance, resistance, or both).

K - Acute/chronic: majority type of physical exercise time.

L – Gender: gender evaluated in the study (M= male; F= female or both).

M - Age range: age range of the study subjects.

N – N: number of the study subjects.

O - Analytical method: microarray, RNA-Seq, genotyping, or qPCR.

P - Up-regulation: positive genetic expression regulation, increased gene expression.

Q - Down-regulation: negative genetic expression regulation, decreased gene expression.

R - Fold change: ratio measures change in the expression level of a gene.

S – FDR: False Discovery Rate: ratio measures the statistical significance.

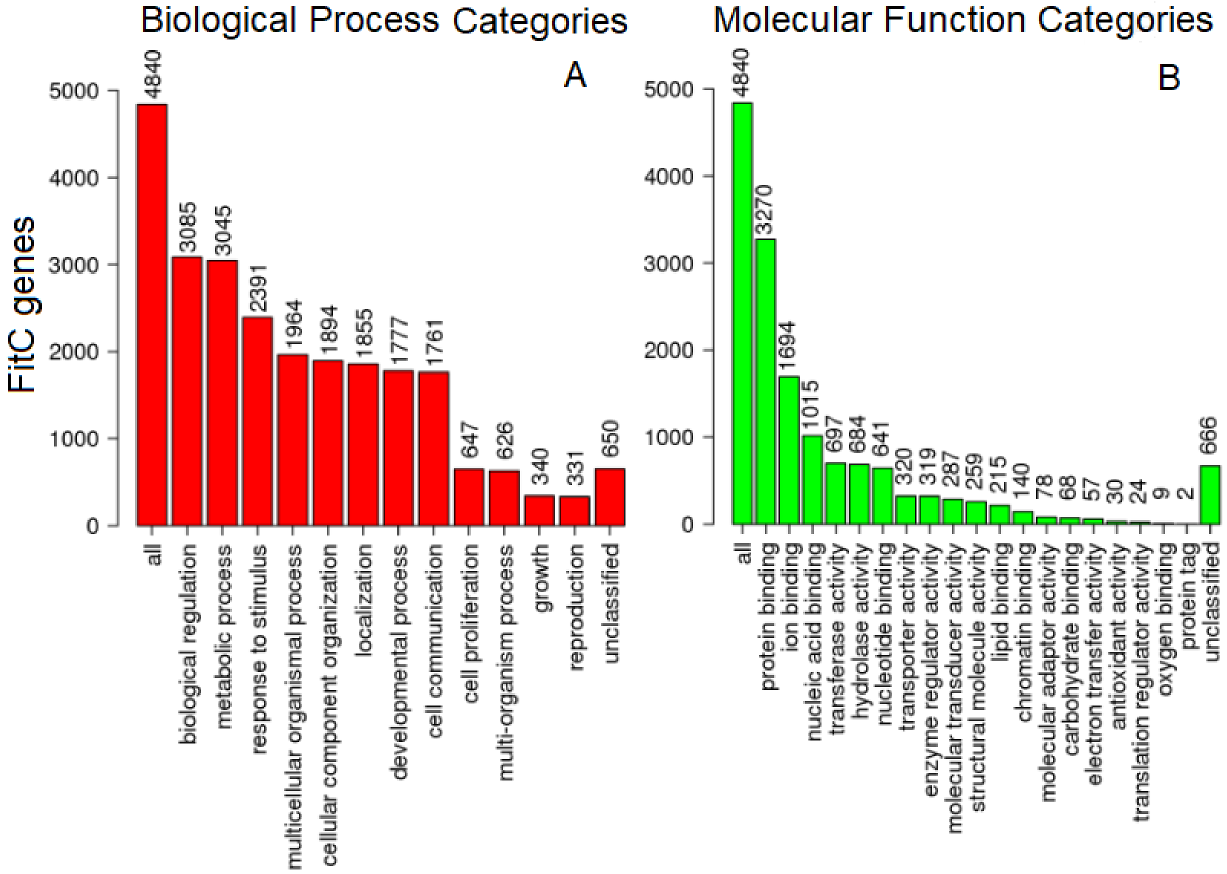
T - p-value: the evidence that you should reject the null hypothesis.

U – References: Used gene references in a short presentation.

Fitnome Catalog References column field

A - This tab presents the used studies' full references, showing 58 lines corresponding to 58 references.

Figure 1. Data functional hierarchization of the FitC dataset based on enrichment of 4,840 genes of FitC by Over Representation Analysis (ORA) method. The height of the bar represents the number of IDs in the user list and the category. A – Biological process categories. B – Molecular Function Categories.



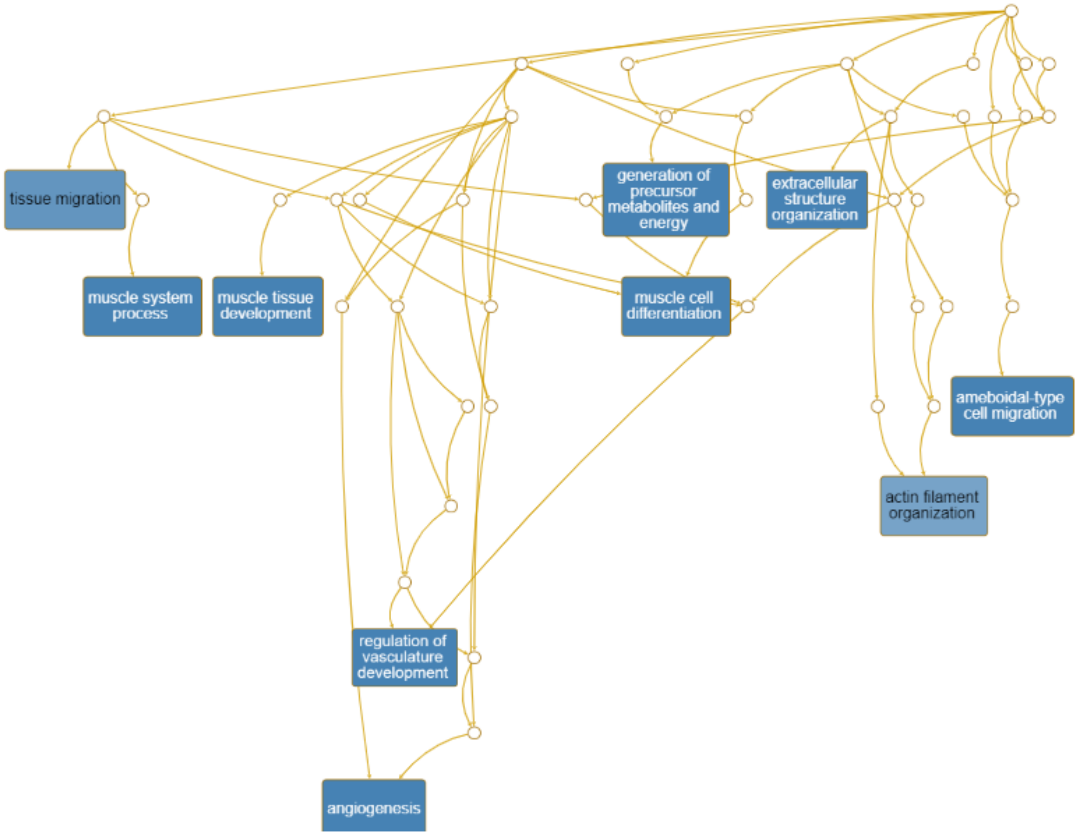
SUPPLEMENTARY MATERIALS

Dataset: "A compendium of physical exercise-related human genes: an 'omic scale analysis."

ACKNOWLEDGEMENTS

This study was sponsored by the National Scientific and Technological Development Council (CNPq) and the Cearense Foundation of Scientific and Technological Support (FUNCAP – CE).

Figure 2. Data functional hierarchization of the FitC gene dataset. This Directed Acyclic Graph ontology process (DAG) was based on showing canonical metabolic pathways. GO terms were shown in the blue rectangles.


tissue migration
muscle system process
muscle tissue development
generation of precursor metabolites and energy
extracellular structure organization
muscle cell differentiation
ameboidal-type cell migration
actin filament organization
regulation of vasculature development
angiogenesis

REFERENCES

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), v. 57, n. 1, p. 289–300, 11 out. 1995.

BOYLE, E. I. et al. GO::TermFinder - Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics, v. 20, n. 18, p. 3710–3715, 2004. DOI: https://doi.org/10.1093/bioinformatics/bth456.

COLBERG, S. R. et al. Exercise and Type 2 Diabetes: The American College of Sports Medicine and the American Diabetes Association: Joint Position Statement. Diabetes Care, v. 33, n. 12, p. e147-67, Dez. 2010. DOI: https://doi.org/10.2337/dc10-9990.

FONTAINE, J. F. et al. Génie: Literature-based gene prioritization at multi genomic scale. Nucleic Acids Research, v. 39, n. SUPPL. 2, p. 455–461, 2011. DOI: https://doi.org/10.1093/nar/gkr246.

GOMES, C. P. C. et al. Chapter Three - Omics and the molecular exercise physiology. In: MAKOWSKI, Gregory S B T - Advances in Clinical Chemistry (Org.). [S.l.]: Elsevier, v. 96. p. 55–84. 2020. DOI: https://doi.org/10.1016/bs.acc.2019.11.003.

KINSELLA, R. J et al. Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space. Database: The Journal of Biological Databases and Curation, v. 2011, p. bar030, 2011. DOI: https://doi.org/10.1093/database/bar030.

LIAO, Y. et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research, v. 47, n. W1, p. W199–W205, 2019. DOI: https://doi.org/10.1093/nar/gkz401.

MAGLOTT, D. et al. Entrez Gene: Gene-Centered Information at NCBI. Nucleic Acids Research, v. 35, n. Database issue, p. D26-31, Jan. 2007.

ORTON, R. J. et al. Bioinformatics tools for analyzing viral genomic data. OIE Revue Scientifique et Technique, v. 35, n. 1, p. 271–285, 2016. DOI: 10.20506/rst.35.1.2432.

PACHECO, C. et al. A compendium of physical exercise-related human genes: An' omic scale analysis. Biology of Sport, v. 35, n. 1, 2018. DOI: 10.5114/biolsport.2018.70746.

SEO, D. Y. et al. Cardiac adaptation to exercise training in health and disease. Pflugers Archiv European Journal of Physiology, exercise in animals - Revision with a table of works, v. 472, n. 2, p. 155–168, 2020.

STRÖHLE, A. Physical Activity, Exercise, Depression and Anxiety Disorders. Journal of Neural Transmission (Vienna, Austria : 1996), v. 116, n. 6, p. 777–784, Jun. 2009.

VASCONCELOS-FILHO, F. S. L. et al. Effect of Involuntary Chronic Physical Exercise on Beta-Amyloid Protein in Experimental Models of Alzheimer's Disease: Systematic Review and Meta-Analysis. Experimental Gerontology, v. 153, p. 111502, out. 2021. DOI: https://doi.org/10.1016/j.exger.2021.111502.

VOLTARELLI, V. A.; FERNANDES, L. G.; BRUM, P. C. Cellular and molecular exercise physiology. Revista Brasileira de Educação Física e Esporte, v. 34, n. 3, p. 533–542, 2020. DOI: https://doi.org/10.11606/1807-5509202000030533.